OpenMPR: Recognize Places Using Multimodal Data for People with Visual Impairments

Ruiqi Cheng & Kaiwei Wang & Jian Bai

State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, Hangzhou, China

E-mail: wangkaiwei@zju.edu.cn

Zhijie Xu

School of Computing and Engineering, University of Huddersfield, Queensgate, Huddersfield, UK

November 2018

Abstract. Place recognition plays a crucial role in navigational assistance, and is also a challenging issue of assistive technology. The place recognition is prone to erroneous localization owing to various changes between database and query images. Aiming at the wearable assistive device for visually impaired people, we propose an open-sourced place recognition algorithm OpenMPR, which utilizes the multimodal data to address the challenging issues of place recognition. Compared with conventional place recognition, the proposed OpenMPR not only leverages multiple effective descriptors, but also assigns different weights to those descriptors in image matching. Incorporating GNSS data into the algorithm, the cone-based sequence searching is used for robust place recognition. The experiments illustrate that the proposed algorithm manages to solve the place recognition issue in the real-world scenarios and surpass the state-of-the-art algorithms in terms of assistive navigation performance. On the real-world testing dataset, the online OpenMPR achieves 88.7% precision at 100% recall without illumination changes, and achieves 57.8% precision at 99.3% recall with illumination changes. The OpenMPR is available at https://github.com/chengricky/OpenMultiPR.

Keywords: Visual Localization, Computer Vision, Navigational Assistance, Assistive Technology

Submitted to: *Meas. Sci. Technol.*

1. Introduction

Vision provides people with the majority of environmental information. Up to 253 million people in the world are with visual impairments (Bourne, Flaxman, Braithwaite, Cicinelli, Das, Jonas, Keeffe, Kempen, Leasher, Limburg, Naidoo, Pesudovs, Resnikoff, Silvester, Stevens, Tahhan, Wong, Taylor, Bourne, Ackland, Arditi, Barkana, Bozkurt, BRAITHWAITE, Bron, Budenz, Cai, Casson, Chakravarthy, Choi, Cicinelli, Congdon, Dana, Dandona, Dandona, Das, Dekaris, Monte, Deva, Dreer, Ellwein, Frazier, Frick, Friedman, Furtado, Gao, Gazzard, George, Gichuhi, Gonzalez, Hammond, Hartnett, He, Hejtmancik, Hirai, Huang, Ingram, Javitt, Jonas, Joslin, Keeffe, Kempen, Khairallah, Khanna, Kim, Lambrou, Lansingh, Lanzetta, Leasher, Lim, LIMBURG, Mansouri, Mathew, Morse, Munoz, Musch, Naidoo, Nangia, PALAIOU, Parodi, Pena, Pesudovs, Peto, Quigley, Raju, Ramulu, Resnikoff, Robin, Rossetti, Saaddine, SANDAR, Serle, Shen, Shetty, Sieving, Silva, Silvester, Sitorus, Stambolian, Stevens, Taylor, Tejedor, Tielsch, Tsilimbaris, van Meurs, Varma, Virgili, Volmink, Wang, Wang, West, Wiedemann, Wong, Wormald & Zheng 2017), and they encounter various difficulties in their daily life. The visually impaired people have limited capability to acquire spatial knowledge (Schinazi, Thrash & Chebat 2016), hence visual place recognition is desired by the visually impaired people, especially in the complex and unfamiliar outdoor environments.

Among the decades, GNSS (global navigation satellite system) has become a prevailing approach to positioning in many applications, such as vehicle navigation, engineering measurement and etc. In order to promote the positioning performance, a number of GNSS processing methods (Paziewski, Sieradzki & Baryla 2018, Odolinski & Teunissen 2017, Odolinski, Teunissen & Odijk 2015, Guo & Zhang 2014, Realini & Reguzzoni 2013) were proposed by the research community to reduce the localization error up to even several millimeters. However, on the low-cost portable devices, the performance of GNSS localization is usually insufficient for the localization demands of the visually impaired people. Compared with that, optical images containing extra positioning cues could be exploited to achieve precise localization. Leveraging images to localize is known as *place recognition*, which is to select the corresponding image of a given query image from database images.

The challenging issues of place recognition lie in applying the place recognition algorithm to real-world scenarios, where the visual appearance of query and database images suffers from variations, such as illuminance changes and viewpoint changes (Lin, Cheng, Wang & Yang 2018). With the proliferation of computer vision, the challenging place recognition task has attracted many researchers to make contributions in this area. Apart from the appearance changes between database and query images, the navigational assistance for people with visual impairments brings in more challenges for the task of place recognition. In the research area of intelligent vehicles, the stationary car-mounted cameras capture the images with high resolution and large field of view, and the accuracy of around several tens of meters is sufficient for car localization. However, the images captured by the wearable devices usually feature low quality, such as the severe motion blur and the continuously changing viewpoint. Moreover, assistive navigation requires more accurate localization, especially at some key positions like street corners, gates and bus stations.

In our previous work (Cheng, Wang, Lin & Yang 2018), multimodal images and GNSS data were used to achieve key position prediction, which aimed to localize the visually impaired person at the positions of interest. Besides, we also implemented Visual Localizer (Lin et al. 2018), which utilized CNN (convolutional neural network) descriptor and data association graph to achieve place recognition for visually impaired people. Aiming at the scenarios of assistive technology, we propose a real-time place recognition algorithm OpenMPR (open-source multimodal place recognition), which extends our preceding research. In this paper, the multiple descriptors of multimodal data and parameter tuning schemes are incorporated to robustify the performance of place recognition in real world. Compared with existing algorithms, OpenMPR runs in an online fashion that only the "past" query images are utilized for place recognition, hence it could be used on wearable assistive devices in real time.

The place recognition procedures of OpenMPR are shown in Figure 1. Multiple descriptors are extracted from multimodal data in both database and query sequences, and the multiple distance matrices are subsequently calculated. Subsequently, the score matrix is synthesized by the distance matrices of different modal data. Finally, the place recognition results are selected from the candidates with high matching scores. The contributions of this paper are summarized as follows:

- To cope with the appearance changes in place recognition, multimodal data, including the images of different modalities and GNSS data, are leveraged for place recognition tasks.
- In order to exploit the latent "place fingerprint" embedded in those data, training-free multiple image descriptors are utilized. The weights of those descriptors are tuned to improve the performance of place recognition.

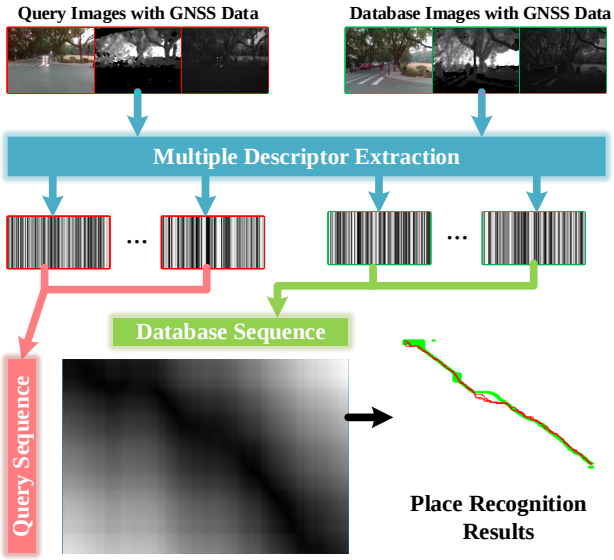

Figure 1. The schematic diagram of OpenMPR, an open-source multimodal place recognition algorithm proposed in this paper.

- Aiming at tackling the localization issues of people with impaired vision, we propose an online place recognition algorithm OpenMPR that surpasses the state of the art and a place recognition dataset collected in the real-world scenarios.

This paper is organized as follows. The related work on place recognition is described briefly in Section 2. The place recognition algorithm based on multimodal data implemented in OpenMPR is presented in Section 3. Moreover, the comprehensive performance experiments are detailed in the Section 4. The last section concludes the paper and presents future work.

2. State of the Art

Place recognition is a prevalent research topic among the communities of computer vision and robotics. According to the types of map abstraction, visual localization falls into metric place recognition and topological place recognition (Lowry, Sünderhauf, Newman, Leonard, Cox, Corke & Milford 2016). Metric place recognition returns localization results with metric information. It includes various SLAM (simultaneous localization and mapping) systems (e.g. ORB-SLAM2 (Mur-Artal & Tardos 2017)) and deep pose prediction networks (e.g. PoseNet (Kendall, Grimes & Cipolla 2015)). Although SLAM systems build the three-dimensional metric maps which could be reused to estimate precise camera poses, they are not suitable for visual localization in changing and large-scale outdoor environments. The deep networks though feature superior robustness against appearance changes, they need to be trained exclusively for each region to predict camera poses in that specific region. For building metric maps, video streams are required as input data to ensure enough scene overlap between successive frames, which is not necessarily available to the wearable assistive devices with limited computational resources. Therefore, metric place recognition is not the optimal choice for assistive technology. Avoiding to build metric maps, topological place recognition generates localization results without metric information. Topological place recognition is suitable for assistive navigation, considering it does not require high-performance hardware and ideal environments.

The community of autonomous vehicles has developed numbers of algorithms to pursue better performance on topological place recognition. Using bag-of-word method, OpenFAB-MAP (Glover, Maddern, Warren, Reid, Milford & Wyeth 2012) is one of earliest open-source packages to achieve appearance-based place recognition. Different kinds of data were leveraged in the existing place recognition algorithms. GNSS priors were exploited in the computationally expensive matching process based on minimum network flow model (Vysotska, Naseer, Spinello, Burgard & Stachniss 2015). Sequence-based LDB (local difference binary) features derived from intensity, gradient and disparity images were utilized to depict images and achieved life-long visual localization in OpenABLE (Arroyo, Alcantarilla, Bergasa & Romera 2016). However, the multimodal LDB descriptors were simply concatenated into single image feature, thus the weights of different modalities in place recognition were not considered. Multiple descriptors were leveraged to achieve sequence-based image matching (Han, Wang, Huang & Zhang 2018), but only color images were used as visual knowledge. Taking advantage of sequence search and match selection, OpenSeqSLAM2.0 (Talbot, Garg & Milford 2018) designed configurable parameters to explore the optimal performance of place recognition under changing conditions.

The appearance variations impede the performance of visual place recognition, and many researchers are dedicated to mitigating the impact of appearance variations towards place recognition by different methods (Lin et al. 2018, Cheng et al. 2018, Kendall et al. 2015, Arandjelović, Gronat, Torii, Pajdla & Sivic 2018). The illumination change is one of vital appearance variations, and quite a few place recognition algorithms (Maddern, Stewart, McManus, Upcroft, Churchill & Newman 2014, Lowry & Milford 2016) addressed the issue. Illumination invariant transformation was proposed to improve visual localization performance during daylight hours (Maddern et al. 2014). Change removal based on unsupervised learning was

utilized to achieve robust place recognition under day-to-night circumstances (Lowry & Milford 2016). Despite the fact that inspiring progress has been obtained by those work, there are challenging issues to be addressed on place recognition for assistive navigation, which has not aroused the sufficient attention of the research community.

To evaluate the performance of place recognition, substantial datasets were proposed by the research community, and some typical datasets feature different appearance variations between query and database images. Those datasets involve the cross-season Nordland dataset (Sünderhauf, Neubert & Protzel 2013) as well as Gardens Point Walking dataset (Sünderhauf, Shirazi, Dayoub, Upcroft & Milford 2015) with viewpoint and illuminance variations. Bonn dataset (Vysotska & Stachniss 2017) and Freiburg dataset (Vysotska et al. 2015) both feature multiple variations, including season, illuminance and viewpoint changes. Most of the datasets are designed for place recognition on autonomous vehicles, the images captured by car-mounted cameras are different with those captured by wearable devices. Besides, the ground truths of those datasets are labeled with GNSS data, hence the localization resolution is not sufficient for assistive technology. To the best of our knowledge, the dataset with multimodal images for assistive technology has not been released.

3. OpenMPR

Different from the existing place recognition approaches, OpenMPR leverages multimodal data to address the issues of place recognition. Apart from vanilla color images, other visual modalities (i.e. depth images and infrared images), as well as GNSS data, are also considered in the system. Multiple descriptors are utilized to exploit the latent information embedded in the multimodal images. The distance matrices derived from the multiple descriptors of query and database images are merged into a synthetic score matrix. Subsequently, the sequence-based matching and selection are executed to obtain the final place recognition results.

3.1. Multiple descriptors extraction from multimodal images

The multimodal images involved in OpenMPR are color images, depth images and near-infrared images. The vanilla color image is an indispensable modality in place recognition task, in that it conveys the both holistic scenes and local textures with chromatic visual cues. Compared with color images, infrared images occupy a longer-wavelength band in spectrum, thus naturally carry different scene information. Depth images contain the three-dimensional shapes, which reduce the odds of mismatching between query and database images. In order to describe the scenes comprehensively, not only are the multimodal images captured to enrich the input information but also both holistic and local image descriptors are utilized to extract the key visual cues embedded in images. As shown in Figure 2, four training-free and pre-trained descriptors are chosen to depict scenes, which avoids the training procedures toward the regions to be deployed so as to be applied to assistive navigation.

The descriptor vector extracted by descriptor f from modality m is defined as $\mathbf{d}^{f,m}$ in this paper. The descriptor f could be one of descriptors in the set

$$F = \{f|GIST,\ LDB,\ BoW,\ CNN\}, \tag{1}$$

and the modality m could be one of modalities in the set

$$M = \{m|color,\ depth,\ infrared\}. \tag{2}$$

The concrete extraction configurations of those descriptors have been illustrated in our previous work (Cheng et al. 2018, Lin et al. 2018). Herein, we summarize descriptor extraction as follows.

3.1.1. Bag of words Based on the local feature ORB (oriented FAST and rBRIEF) (Rublee, Rabaud, Konolige & Bradski 2011), BoW (bag of words) characterizes the image details by the occurrence of each visual word clustered by local features. BoW is widely applied to object and scene categorization, due to its simplicity, computational efficiency and invariance to affine transformation (Galvez-López & Tardos 2012). In this paper, the key points [see Figure 2 (c1)] are detected by oriented FAST (features from accelerated segment test) and are described by rBRIEF (rotated binary robust independent elementary features). The ORB descriptors of all key points are merged together and compose the concatenated descriptors [see Figure 2 (c2)]. Subsequently, the BoW descriptor [see Figure 2 (c3)] is generated using the extracted ORB descriptors and the pre-trained vocabularies (Muñoz-Salinas 2017). In view that the off-the-shelf vocabularies were trained on photometric images, BoW desceiptors are extracted from color and infrared modalities.

3.1.2. Local difference binary The holistic image descriptors, i.e. GIST (Oliva & Torralba 2001, Torralba, Murphy, Freeman & Rubin 2003), LDB (Yang & Cheng 2014) and CNN descriptors, emphasize whole visual features rather than local details, hence are used to alleviate the impact of appearance changes for image matching. As shown in Figure 2 (a), LDB descriptor is extracted as a global descriptor after the preprocessing

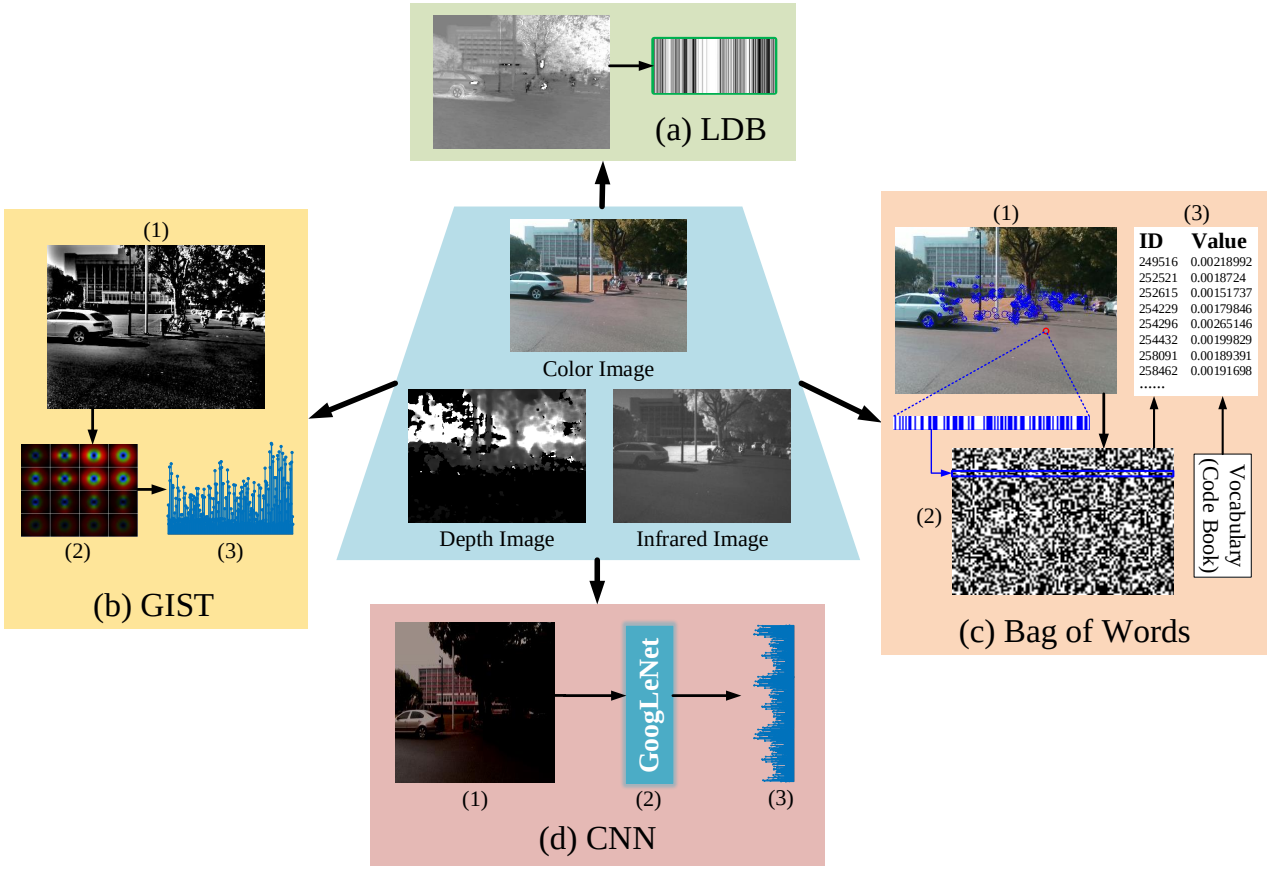

Figure 2. The multiple descriptors extracted from the multimodal images.

of illumination invariance transformation. It is worthwhile to note that bit selection (Yang & Cheng 2014) is not executed in this paper, in that the compression of the global descriptor hinders the performance of image description. LDB descriptors are extracted from all of the modalities separately.

3.1.3. GIST Also as a holistic image descriptor, GIST represents the scene in a very low dimensions. The global descriptor GIST is extracted from the preprocessed image, which involves image normalization [see Figure 2 (b1)], Gabor filtering [see Figure 2 (b2)] and response averaging [see Figure 2 (b3)]. GIST descriptors are extracted from all of the modalities separately.

3.1.4. CNN Different from the hand-crafted descriptors above, the descriptors selected from CNN are also used to enhance the description ability of the system. As presented in Figure 2 (d), the CNN descriptor is generated from the intermediate layers of the pre-trained GoogLeNet fed by the preprocessed color image. The compressed concatenation of two layers $inception3a/3 \times 3$ and $inception3a/3 \times 3\_reduce$ in GoogLeNet pre-trained on Places365 dataset (Zhou, Lapedriza, Khosla, Oliva & Torralba 2017) is used for image descriptor. Constrained by the structure of GoogLeNet, the CNN descrtiptor is only extracted from color images.

3.2. Distance matrices with GNSS priors

The extracted multiple descriptors $\{\mathbf{d}^{f,m}|f \in F, m \in M\}$ are leveraged to measure the similarity between images, thus to characterize the correspondence of query images and database images. In this paper, sequential images, rather than single images, are utilized during image matching. Assuming that the query sequence has the size of n and the database sequence has the size of l, then the distance matrix features the size of $n \times l$. Herein, we define $D^{f,m}$ as the distance matrix of descriptor f extracted from modality m. The element $D^{f,m}_{i,j}$ of the matrix $D^{f,m}$ is attained by measuring the descriptor distance between the i-th query image and the j-th database image. The distance measurement varies from different descriptors. For binary descriptors (LDB), Hamming distance is measured as the distance of images, while the distances

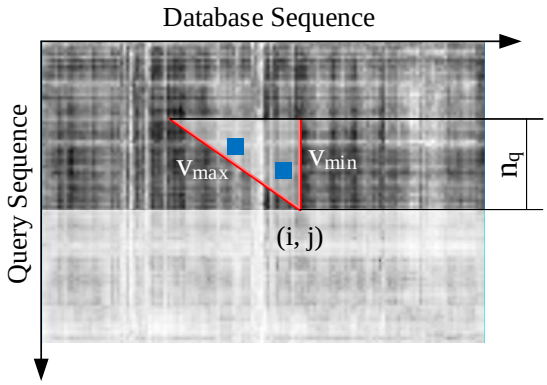

Figure 3. The online cone-based searching schematics.

of GIST and CNN descriptors are measured with Euclidean distance.

Despite with the insufficient positioning accuracy, GNSS data consisting the coordinates of longitude and latitude provide with a priori knowledge for visual place recognition. With the GNSS priors, those query-database pairs that leave a large spatial distance between each other need not be matched, so as to improve the computational efficiency and to reduce the possibility of image mismatching. The metric distance between the i-th query image and the j-th database image is specified as $G_{i,j}$, hence the final distance matrix containing GNSS data $E^{f,m}$ is obtained by

$$E_{i,j}^{f,m} = \begin{cases} D_{i,j}^{f,m} & G_{i,j} \leq g, \\ Inf. & G_{i,j} > g. \end{cases}, \tag{3}$$

where g is the threshold of possible matching pairs. The smaller the threshold g, the smaller the searching range of image matching. Considering the observation error of the GNSS module used in this paper, the threshold g should not be too small, the correct matching results would be ruled out otherwise. In this paper, g is set to 15 meters.

3.3. Online sequence-base searching and matching scoring

Having obtained a distance matrix $E^{f,m}$, we execute an online cone-based searching upon every query-database pair, which achieves sequential image matching and gets a matching score for each pair.

As shown in Figure 3, the horizontal axis denotes the database sequence, and the vertical axis denotes the query sequence. Within the distance matrix, each query-database pair (i, j) is associated with only one cone region which is limited by sequential length n_q, maximal velocity v_{max} and minimal velocity v_{min}. The online cone-based searching algorithm proposed in this paper is different from the offline one in (Milford, Firn, Beattie, Jacobson, Pepperell, Mason, Kimlin & Dunbabin 2014). The offline searching algorithm makes use of the "future" query images, thus place recognition cannot run in real time.

Within the region, the number of best-matching pairs (represented by blue squares in Figure 3) is counted firstly. The best-matching pair is defined as the minimum value of a certain row in the distance matrix. In other words, a query descriptor and the database descriptor featuring minimum distance with that query descriptor compose a best-matching pair. Herein, the number of best-matching pairs in a cone region is defined as n_{match}, and the score $s_{i,j}$ of the query-database pair (i, j) is defined as

$$s_{i,j} = \frac{n_{match}}{n_q}. \tag{4}$$

Naturally, all of the matching scores $s_{i,j}$ form into a score matrix $S^{f,m}$. Multiple descriptors extracted from different modalities carry diverse visual information, so assigning the same weight to different descriptors during image matching does not necessarily promote the matching robustness. Therefore, the coefficients of score matrix synthesis $\{\lambda^{f,m}\}$ need to be adjusted for the better accuracy of place recognition. The score matrices derived from different descriptors of different modalities are synthesized to a single score matrix S, which is presented as

$$S_{i,j} = \frac{\sum\limits_{f \in F, m \in M} \lambda^{f,m} \times S_{i,j}^{f,m}}{\sum\limits_{f \in F, m \in M} \lambda^{f,m}}. \tag{5}$$

The genetic algorithm (Mohammadi, Asadi, Mohamed, Nelson & Nahavandi 2017) is used to determine the values of $\{\lambda^{f,m}\}$, which is described later in Section 4. With matching score matrix, each query image corresponds to the best database image with the highest score. In order to get the final place recognition results, the matching score of the best query-database pair is evaluated to rule out the mismatching pairs. In this paper, we use score thresholding (Talbot et al. 2018) to remove low-confidence matching results whose score lower than threshold t.

3.4. Implementation

The proposed OpenMPR algorithm is implemented in C++, considering the portability and effectiveness. The open-source code of OpenMPR is available online (Cheng 2019). The dependencies include *OpenCV 4.0* (OpenCV 2018), as well as *DBoW3* (Muñoz-Salinas 2017) for BoW extraction, *LibGIST* (Song 2014) for GIST extraction, *libLDB* (Yang & Cheng 2014) for LDB extraction, and *OpenGA* (Mohammadi et al. 2017) for parameter tuning.

The settings of OpenMPR could be easily switched by configuration file *Config.yaml*. There are two modes

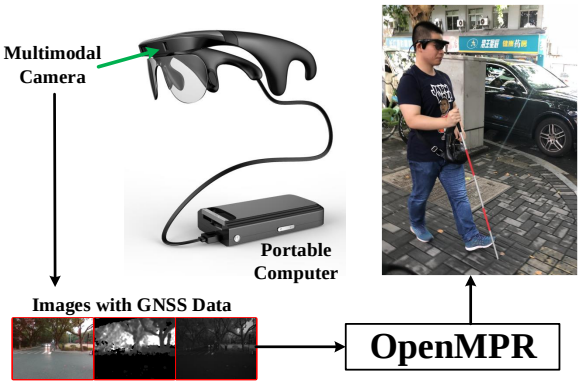

Figure 4. The assistive device Intoer is used to capture multimodal data.

in OpenMPR, which is testing mode and tuning mode. In testing mode, the place recognition is executed by using the default or customized parameters. In tuning mode, the optimal parameters are searched to achieve the best performance. The other configurable parts of OpenMPR involve the resolution of input images, whether to use GNSS data, and whether to use certain image modality or image descriptor.

4. Experiments

In this section, the real-world place recognition dataset collected by the assistive device is illustrated firstly. In order to achieve the optimal performance of place recognition, the experiments on parameter tuning were carried out and the tuning results are analyzed thoroughly. Finally, the state-of-the-art performance of OpenMPR is validated through comparative study.

In view that OpenMPR is prone to be implanted into assistive devices, the experiments were carried out on the assistive device Intoer (KrVision 2017), which is shown in Figure 4. The assistive device Intoer is utilized not only to capture multimodal images and GNSS data but also to run the OpenMPR algorithm.

4.1. Datasets

In view that the place recognition dataset with multimodal data has not been released, we collected a real-world dataset, available at (Cheng 2019), within the campus of Yuquan, Zhejiang University.

One frame of data consists of a color image, a depth image, an infrared image, and a GNSS coordinate. The multimodal images were collected using Intel RealSense ZR300 Camera (Keselman, Woodfill, Grunnet-Jepsen & Bhowmik 2017) embedded in Intoer, which is an infrared assisted stereo vision camera. In terms of the effective range and density of depth images, Realsense ZR300 represents the moderate level among commercial RGB-D cameras. Thereby, the dataset proposed in this paper is adequate to evaluate the performance of OpenMPR. Other imaging specifications are illustrated in Table 1. The GNSS data were collected with the customized GNSS receiver embedded in Intorer.

Table 1. The specifications of multimodal images captured by ZR300.

	Color	Depth	Infrared
Resolution	320 × 240	320 × 240	320 × 240
Shutter type	Rolling	Global	Global
Frame rate	1 FPS	1 FPS	1 FPS

Up to 1,671 frames of data are involved in the dataset, where the four subsets were collected on three routes as shown in Table 2 and Figure 5. Although Train-1 and Train-2 cover the same route, they were collected in the opposite traversing direction. It is worthwhile to note that no route overlap exists among the training subsets and the testing subsets. Moreover, images of the four subsets are not selected artificially. In the experiments, Train-1 and Train-2 are utilized to tune the parameters, and Test-3 and Test-4 are used to validate the performance of multimodal place recognition.

Each subset is composed of one query sequence and one database sequence. The collected multimodal images feature apparent viewpoint changes between query and database sequence, since the camera is embedded in the wearable device and the query and database were not captured on the completely identical route. Apart from that, all of the images also present dynamic object changes between query and database. For example, the person passing by in front of the camera appears in the query sequence, but does not appear in the database sequence. Moreover, the illumination changes exist in Train-2 and Test-4. In those subsets, database and query images were captured in the afternoon and at dusk separately. All of those real-world changes form into substantial challenges for place recognition. In brief, the dataset was collected in real-world scenarios, and is suitable for the localization issues of assistive technology.

Due to the inaccuracy of GNSS data, the ground truth of dataset is labeled manually according to visual similarity rather than GNSS distance. Each query image is associated with the best-matching database image.

4.2. Parameter tuning procedures and results

In order to optimize the performance of place recognition, a series of parameters are tuned on the training datasets that are separated from the testing datasets. The parameters include the length (n_q) and

Figure 5. The three traversed paths in multimodal datasets of place recognition. Route A: from the teaching building to the gate (orange). Route B: from the teaching building to the library (green). Route C: from the library to the teaching building (blue).

Table 2. The characteristics of multimodal place recognition dataset. v = viewpoint changes, o = dynamic objects, and i = illumination changes.

Subset	Route	# query	# database	Changes
Train-1	C	212	291	v/o
Train-2	C	208	215	v/o/i
Test-3	A	97	142	v/o
Test-4	A+B	233	273	v/o/i

velocity limits (v_{max} and v_{min}) of cone searching, the coefficients of score matrix synthesis ($\lambda^{f,m}$), and the threshold t of score thresholding.

If a query image matches with a database image, the result is defined as a positive result. If no database image is matched, the result is defined as a negative result. Considering the query and database images are sequential in the dataset, the place recognition result of a query image could be represented as the sequential index of the best-matching database image. If the index difference between the place recognition result and the ground truth is less than or equal to the tolerance (set to 5 in this paper), the result is defined as a TP (true positive) result. Otherwise, the positive result is defined as a FP (false positive) result. Moreover, if the result should match with a database image but it does not match with any database image, the result is defined as a FN (false negative) result. The performance of OpenMPR is evaluated and analyzed in terms of precision and recall. Precision is the proportion of true positives out of all predicted positives, and recall is the proportion of true positives to all of actual positives.

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

In this section, the objective of parameter tuning is to choose the parameters that achieve the greatest F_1 score.

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{8}$$

The following section outlines the procedures and results of configurable parameter tuning.

4.2.1. Coefficients of score matrix synthesis The coefficient $\lambda^{f,m}$ denotes the importance of specific descriptor $\mathbf{d}^{f,m}$ during place recognition. In order to tune the coefficients efficiently, we leverage the genetic algorithm implemented by (Mohammadi et al. 2017) to seek the optimal combination of coefficients. The length of the cone region (n_q) is set to 1, in order that the sequential searching does not affect coefficient optimization.

The genetic algorithm is an analogue of natural selection, which optimizes the parameters (genes) by the bio-inspired operators such as mutation, crossover and selection. The coefficient array with the size of 9 (see Table 3) is defined as the genes, and the fitness of genes is evaluated with the F_1 score at that coefficient combination. The principles and implementation details of the genetic algorithm could be found in (Mohammadi et al. 2017). Empirically, the maximum number of generations (80 in this paper) is set as the stopping criterion, which is sufficient for the genetic algorithm to generate a stable iterative result. The genetic algorithm runs for multiple times (15 in this paper) with randomly initialized genes to avoid the local optima. The mean coefficients of the multiple results obtained by the genetic algorithm are set as final parameter searching results. The mean coefficients of the two datasets are chosen as the optimal parameters, as presented in Table 3.

As demonstrated in Table 2, Train-1 suffers from viewpoint changes and dynamic objects, meanwhile Train-2 suffers from more illumination changes than Train-1. On Train-2, the CNN descriptor presents the highest weights compared with other descriptors, which illustrates that the descriptors derived from the GoogLeNet pre-trained on Places365 yield superior description performance even under the severe changes. On dataset 1, the GIST descriptors show better description performance compared with other descriptors, which indicates that GIST descriptor is suitable for depicting the images without large illumination changes. Besides, LDB presents the suboptimal performance of place recognition in the complicated environments. The dataset used in this paper features severe viewpoint changes and dynamic objects, hence the holistic descriptors are important to grasp the global information.

Compared with the holistic descriptors, the performance of BoW descriptors on the two datasets reveals that BoW is advantageous and stable for place

Table 3. The searching results of coefficients $\lambda^{f,m}$ using the genetic algorithm. (c=color, d=depth, i=infrared.)

Dataset	$f = BoW$		$f = GIST$			$f = LDB$			$\lambda^{CNN,c}$	F_1
	$\lambda^{f,c}$	$\lambda^{f,i}$	$\lambda^{f,c}$	$\lambda^{f,d}$	$\lambda^{f,i}$	$\lambda^{f,c}$	$\lambda^{f,d}$	$\lambda^{f,i}$		
Train-1	1.359	1.666	2.269	1.102	0.986	0.617	0.469	1.042	0.491	0.73
Train-2	1.132	1.705	0.889	1.081	0.989	0.436	0.778	0.638	2.353	0.63
Optimal	1.245	1.685	1.579	1.091	0.987	0.526	0.623	0.840	1.422	-

recognition in various environments. More conclusions can be drawn when inspecting the results on color and infrared modality carefully. The BoW descriptor on the infrared modality features a higher weight than that on color modality. The reasonable explanation is that the local ORB features are susceptible to image details with motion blur, which is prone to occur on color images captured with the rolling shutter. On the contrary, the infrared modality features better performance on imaging stability thanks to the global shutter.

4.2.2. Parameters of cone-based searching Having chosen the optimal coefficients (shown in Table 3), the tuning procedures of other parameters are executed. The parameters of cone-based searching algorithm include the length of sequence (n_q) and velocity limits (v_{max} and v_{min}). They represent the quantity of information used in cone searching. In parameter sweeping, the maximal velocity (v_{max}) is set as the reciprocal of the minimal velocity (v_{min}), so there are only two parameters to be tuned. The minimal velocity (v_{min}) is varied from 0.1 to 0.75, and the length (n_q) is varied from 3 to the 79.

Different velocity limits of cone-based searching are utilized to test the performance of place recognition. The Figure 6 demonstrates that $v_{min} \geq 0.4$ ($v_{max} \leq 2.5$) features good performance, and that the larger velocity range results in suboptimal performance. The large searching range introduces more best-matching pairs, meanwhile introduces more potential inaccurate results. The velocity limits should be moderate to tolerant the real-world scenarios, such as the inconsistency of the carrier's walking speed when recording query and database sequences. Thereby, we set $v_{min} = 0.4$, and $v_{max} = 2.5$.

As shown in Figure 7, the performance of place recognition is related to the length of searching sequence (n_q). Whether n_q is too large or too small, the performance is limited. For the sake of computational efficiency, we set the optimal parameter n_q as 10.

4.2.3. Threshold of score thresholding The threshold score thresholding t affects the precision and recall of place recognition. The score threshold t is used to eliminate bad matching results and improve the performance of place recognition. As shown in

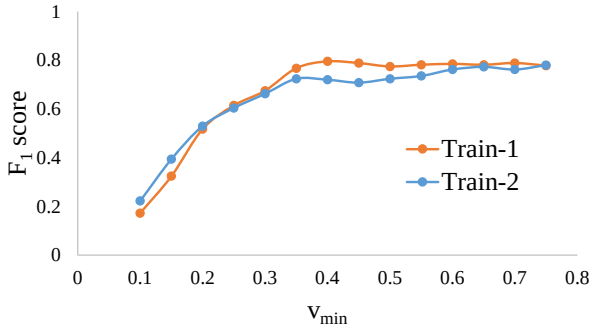

Figure 6. The parameter sweeping results of v_{min}.

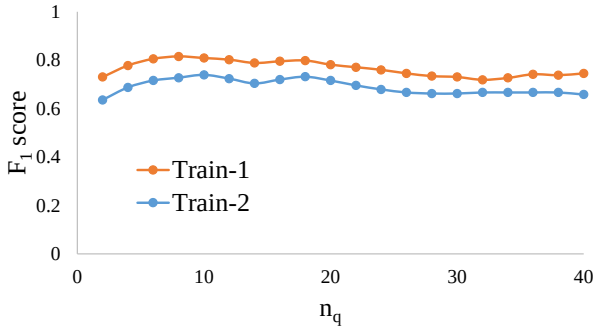

Figure 7. The parameter sweeping results of n_q.

Figure 8, the precision-recall curve under different thresholds t is plotted. As the threshold is low, the matching results with low confidence influence the precision of place recognition. On the contrary, the high threshold results in low recall rate. The optimal value of threshold t is set to 0.16, where the recall has not descended substantially and the precision maintains at a high level.

4.3. Validation of OpenMPR

In order to validate the parameter tuning results and the systematic performance of OpenMPR, the testing sets whose routes are different from those of training sets are utilized to evaluate the proposed algorithm. As demonstrated in Table 2, viewpoint changes and dynamic objects exist in both subsets, meanwhile illumination changes are introduced in Test-4.

4.3.1. Validation of optimized coefficients To validate the effectiveness of optimized coefficients $\{\lambda^{f,m}\}$, the

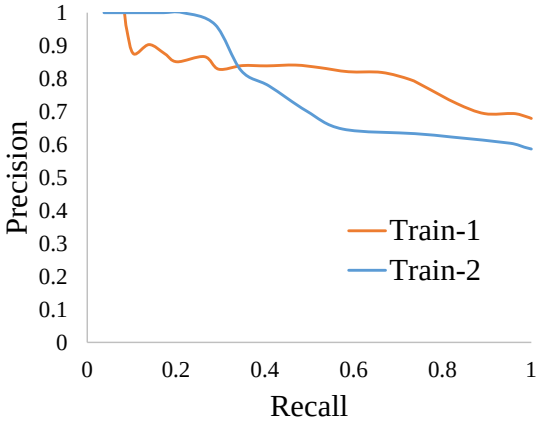

Figure 8. The precision-recall curve as sweeping parameter t.

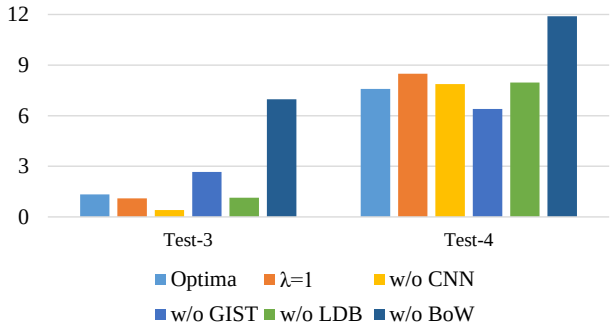

Figure 9. The mean localization error under different configurations of parameters.

place recognition performance on different coefficient configurations is compared. In addition to optimized coefficients, the other configuration involves

(1) $\lambda = 1$: let all of the coefficients be 1, which means that all descriptors feature the same importance.

(2) *w/o certain descriptor*: let the coefficients of the corresponding descriptor be 0, which means abandoning that descriptor in the system.

The mean localization error of different coefficient configurations is shown in Figure 9. Herein, the localization error refers to the index difference between the OpenMPR result and the ground truth. The mean localization error is obtained by averaging the localization error of all query images in the testing set. It is concluded that the configuration of optimized coefficients shows the balanced performance on both testing sets, though it is not the best configuration on the single dataset.

On both testing sets, the configuration *w/o BoW* features the worst performance, which illustrates that BoW descriptor is essential for place recognition. On Test-3, the configuration *w/o CNN* yields the optimal performance, and the configuration *w/o GIST* shows the suboptimal performance. Those phenomena are consistent with the analysis in Section 4.2.1 that the GIST descriptor, instead of the CNN descriptor, plays the vital role in place recognition if there is no illumination changes. In contrast, on the testing set with illumination changes, the GIST descriptor is no longer eligible for good performance, the CNN descriptor and the other descriptors are indispensable.

4.3.2. Validation of systematic performance With the optimal parameters determined in the preceding sections, the place recognition results of OpenMPR on the two testing sets are compared with the state-of-the-art place recognition algorithms. OpenSeqSLAM2.0 (Talbot et al. 2018) and Visual Localizer (Lin et al. 2018) are chosen as the baselines of OpenMPR. Though OpenSeqSLAM2.0 was designed for visual place recognition on autonomous vehicles, it provides with important inspirations in terms of sequence searching and matching selection techniques. In the experiments, the OpenSeqSLAM2.0 parameters related to sequence searching and matching selection were set to those optimal values presented above. As a preliminary work, Visual Localizer proposed a place recognition solution for the mobility of visually impaired people using pretrained CNN descriptors and global optimization.

Table 4. The place recognition results on the testing datasets.

Algorithm	Subset	Precision	Recall	Error
OpenMPR	Test-3	88.7%	100.0%	**1.33**
	Test-4	57.8%	99.3%	**7.59**
OpenSeq-SLAM2.0	Test-3	26.6%	34.0%	7.80
	Test-4	25.7%	82.0%	47.91
Visual Localizer	Test-3	48.5%	100%	19.14
	Test-4	58.4%	100%	9.99

As shown in Table 4, the three performance indicators (precision, recall and mean localization error) are leveraged to evaluate the results of OpenMPR on the two testing sets. In terms of mean localization error, the proposed OpenMPR is superior to the two state-of-the-art algorithms. According to the statistics of OpenMPR, the place recognition on Test-3 is more precise than that on Test-4, in that fewer appearance changes are involved in Test-3. Fortunately, with the help of the multiple descriptors extracted from multimodal images, the mean localization error of OpenMPR on Test-4 is acceptable, which slightly exceeds the tolerance of 5. GNSS priors play an important role in ruling out the definite negatives during image matching.

Compared with OpenMPR, OpenSeqSLAM2.0 yields inferior localization performance both on Test-3 and Test-4, in view of the low recall and large

localization error. Apparently, OpenSeqSLAM2.0 that measures the similarity of images via sum of absolute differences of normalized images does not make place recognition robust against various appearance changes. The comparison between OpenMPR and OpenSeqSLAM justifies that the proposed image descriptors robustify place recognition under practical conditions. For Visual Localizer, the performance of Test-4 with more appearance changes surpasses that of Test-3, which resembles the phenomenon that CNN descriptor features a higher weight on Train-2 than on Train-1. It is evident that the proposed CNN descriptor (the compressed concatenation of $inception3a/3 \times 3$ and $inception3a/3 \times 3\_reduce$) is capable of extracting effective semantic "place fingerprint" between images with large appearance changes. However, without the aid of other descriptors and multimodal images, the performance of the CNN descriptor on Test-3 is limited, which further confirms that the necessity of multiple descriptors proposed in this paper.

As shown in Figure 10, the place recognition result of OpenMPR is visualized as a visualization matrix with the size of $n \times l$, where n is the number of query images and l is the number of database images. The element of the matrix denotes the query-database pair. In the matrix, green and red points denote ground truths (with the tolerance of 5) and localization results respectively. From the diagrams, it is concluded that the place recognition results basically conform to the corresponding ground truths, despite the serious viewpoint variation, motion blur and dynamic objects (e.g. pedestrians). Even on Test-4 with obvious illumination changes, most of the mismatching images are not far from the tolerance of place recognition. In Figure 10, some successful matching results are presented, which indicates that OpenMPR still recognizes places under the conditions of various appearance changes.

Real-time performance is crucial for assistive navigation. OpenSeqSLAM2.0 uses both the "past" images and the "future" images during cone-based searching, hence it cannot be used in real time. Unfortunately, network flow-based global optimization scheme embedded in Visual Localizer features inferior computational efficiency. The single-frame computation speed is analyzed on the Inoter with Intel Atom x5-Z8500 and a desktop with Inter Core i5-6500 to evaluate the real-time performance of OpenMPR, as shown in Table 5. The real-time requirement is basically satisfied by OpenMPR according to the results on the Intoer. With the update of Intoer hardware, the real-time performance would be further improved in view of the speed test on the desktop. After inspecting the running time of descriptor extraction, it is found that time consumed during extracting GIST descriptors from multimodal images accounts for the major proportion (more than 80% of descriptor extraction). In the future, applying GIST library with superior computational efficiency to OpenMPR leads to better real-time performance of the system.

Table 5. The real-time performance of OpenMPR on different platforms.

Platform	Descriptor Extraction	Matching	Overall
Intel Atom x5-Z8500	2,056 ms	98 ms	2,154 ms
Intel Core i5-6500	362 ms	25 ms	387 ms

5. Conclusion

Different with the majority of place recognition work, this paper focus on the traveling demands of visually impaired people, and propose an open-source software OpenMPR, which leverages multi-modal data for online place recognition task.

In the area of assistive technology, the wearable camera tends to capture images with motion blur and low resolution. Due to the limited computational resource, discrete images (one image per second in this paper), instead of video streams, are captured and processed on the portable devices. Apart from that, the query and database sequences features various appearance changes, including viewpoint changes, illumination changes and dynamic objects. In those real-world scenarios, the proposed OpenMPR utilizes configured multiple descriptors extracted from multimodal data and online sequence-based searching to obtain good place recognition performance. It achieves 88.7% precision at 100% recall without illumination changes, and achieves 57.8% precision at 99.3% recall with illumination changes.

In the future, we plan to achieve semantic place recognition, where the visual information in images is understood and places are autonomously labeled with different levels of importance.

6. Acknowledgments

This work was supported by the State Key Laboratory of Modern Optical Instrumentation.

References

Arandjelović, R., Gronat, P., Torii, A., Pajdla, T. & Sivic, J. (2018). NetVLAD: CNN architecture for weakly supervised place recognition, *IEEE*

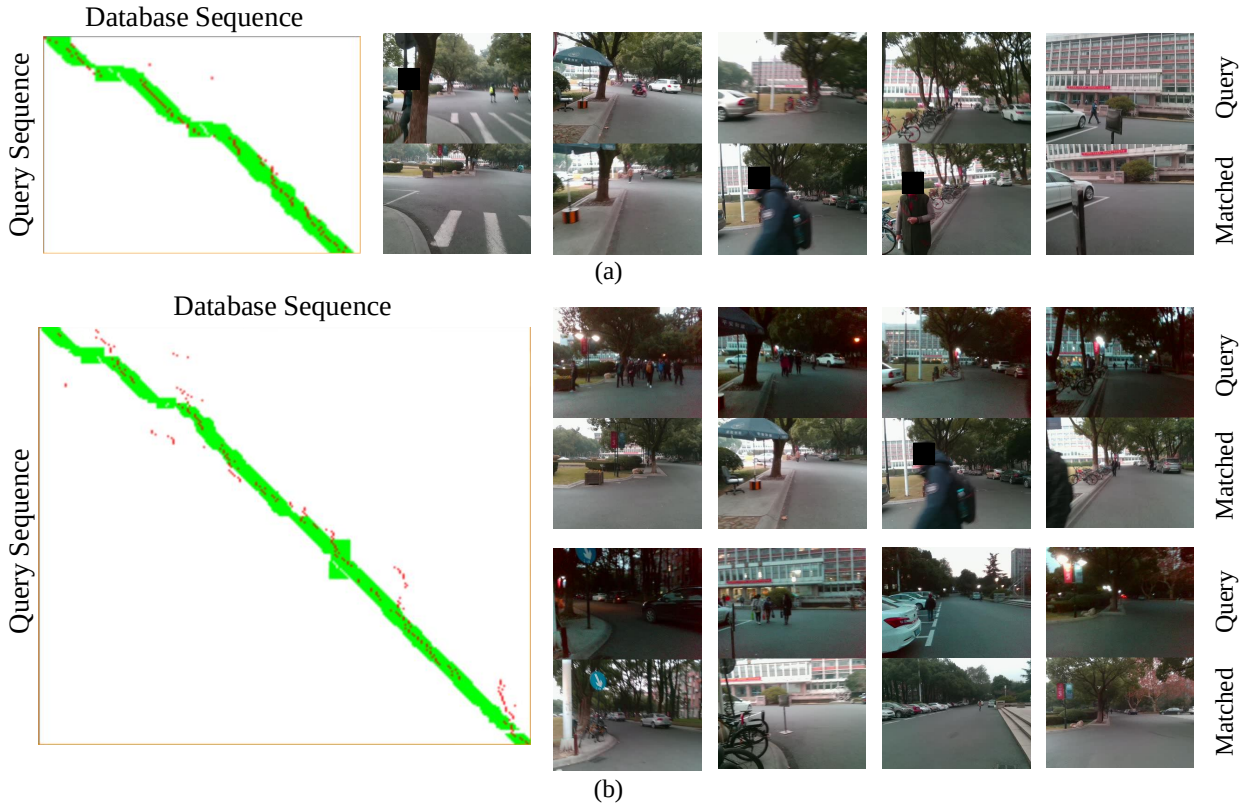

Figure 10. The place recognition results and some localization instances on (a) Test-3 and (b) Test-4 dataset. In the left diagram, the horizontal axis denotes the database sequence, and the vertical axis denotes the query sequence.

Transactions on Pattern Analysis and Machine Intelligence **40**(6): 1437–1451.

Arroyo, R., Alcantarilla, P. F., Bergasa, L. M. & Romera, E. (2016). OpenABLE: An open-source toolbox for application in life-long visual localization of autonomous vehicles, *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 965–970.

Bourne, R. R. A., Flaxman, S. R., Braithwaite, T., Cicinelli, M. V., Das, A., Jonas, J. B., Keeffe, J., Kempen, J. H., Leasher, J., Limburg, H., Naidoo, K., Pesudovs, K., Resnikoff, S., Silvester, A., Stevens, G. A., Tahhan, N., Wong, T. Y., Taylor, H. R., Bourne, R., Ackland, P., Arditi, A., Barkana, Y., Bozkurt, B., BRAITHWAITE, T., Bron, A., Budenz, D., Cai, F., Casson, R., Chakravarthy, U., Choi, J., Cicinelli, M. V., Congdon, N., Dana, R., Dandona, R., Dandona, L., Das, A., Dekaris, I., Monte, M. D., Deva, J., Dreer, L., Ellwein, L., Frazier, M., Frick, K., Friedman, D., Furtado, J., Gao, H., Gazzard, G., George, R., Gichuhi, S., Gonzalez, V., Hammond, B., Hartnett, M. E., He, M., Hejtmancik, J., Hirai, F., Huang, J., Ingram, A., Javitt, J., Jonas, J., Joslin, C., Keeffe, J., Kempen, J., Khairallah, M., Khanna, R., Kim, J., Lambrou, G., Lansingh, V. C., Lanzetta, P., Leasher, J., Lim, J., LIMBURG, H., Mansouri, K., Mathew, A., Morse, A., Munoz, B., Musch, D., Naidoo, K., Nangia, V., PALAIOU, M., Parodi, M. B., Pena, F. Y., Pesudovs, K., Peto, T., Quigley, H., Raju, M., Ramulu, P., Resnikoff, S., Robin, A., Rossetti, L., Saaddine, J., SANDAR, M., Serle, J., Shen, T., Shetty, R., Sieving, P., Silva, J. C., Silvester, A., Sitorus, R. S., Stambolian, D., Stevens, G., Taylor, H., Tejedor, J., Tielsch, J., Tsilimbaris, M., van Meurs, J., Varma, R., Virgili, G., Volmink, J., Wang, Y. X., Wang, N.-L., West, S., Wiedemann, P., Wong, T., Wormald, R. & Zheng, Y. (2017). Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis, *The Lancet Global Health* **5**(9): e888 – e897.

Cheng, R. (2019). OpenMPR.
URL: *https://github.com/chengricky/OpenMultiPR*

Cheng, R., Wang, K., Lin, L. & Yang, K. (2018). Visual localization of key positions for visually impaired people, *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2893–2898.

Galvez-López, D. & Tardos, J. D. (2012). Bags of binary words for fast place recognition in image sequences, *IEEE Transactions on Robotics* **28**(5): 1188–1197.

Glover, A., Maddern, W., Warren, M., Reid, S., Milford, M. & Wyeth, G. (2012). OpenFABMAP: An open source toolbox for appearance-based loop closure detection, *2012 IEEE International Conference on Robotics and Automation*, pp. 4730–4735.

Guo, F. & Zhang, X. (2014). Adaptive robust kalman filtering for precise point positioning, *Measurement Science and Technology* **25**(10): 105011.

Han, F., Wang, H., Huang, G. & Zhang, H. (2018). Sequence-based sparse optimization methods for

long-term loop closure detection in visual slam, *Autonomous Robots* **42**(7): 1323–1335.

Kendall, A., Grimes, M. & Cipolla, R. (2015). PoseNet: A convolutional network for real-time 6-dof camera relocalization, *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2938–2946.

Keselman, L., Woodfill, J. I., Grunnet-Jepsen, A. & Bhowmik, A. (2017). Intel(R) RealSense(TM) stereoscopic depth cameras, *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1267–1276.

KrVision (2017). Intoer: auxiliary glasses for people with visual impairments (in chinese).
URL: *http://www.krvision.cn/cpjs/*

Lin, S., Cheng, R., Wang, K. & Yang, K. (2018). Visual Localizer: Outdoor localization based on ConvNet descriptor and global optimization for visually impaired pedestrians, *Sensors* **18**(8).

Lowry, S. & Milford, M. J. (2016). Supervised and unsupervised linear learning techniques for visual place recognition in changing environments, *IEEE Transactions on Robotics* **32**(3): 600–613.

Lowry, S., Sünderhauf, N., Newman, P., Leonard, J. J., Cox, D., Corke, P. & Milford, M. J. (2016). Visual place recognition: A survey, *IEEE Transactions on Robotics* **32**(1): 1–19.

Maddern, W., Stewart, A., McManus, C., Upcroft, B., Churchill, W. & Newman, P. (2014). Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles, *Proceedings of the Visual Place Recognition in Changing Environments Workshop, IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China*, Vol. 2, p. 3.

Milford, M., Firn, J., Beattie, J., Jacobson, A., Pepperell, E., Mason, E., Kimlin, M. & Dunbabin, M. (2014). Automated sensory data alignment for environmental and epidermal change monitoring, *Australasian Conference on Robotics and Automation 2014*, Australian Robotic and Automation Association, The University of Melbourne, Victoria, Australia, pp. 1–10.

Mohammadi, A., Asadi, H., Mohamed, S., Nelson, K. & Nahavandi, S. (2017). OpenGA, a C++ genetic algorithm library, *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2051–2056.

Mur-Artal, R. & Tardos, J. D. (2017). ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras, *IEEE Transactions on Robotics* **33**(5): 1255–1262.

Muñoz-Salinas, R. (2017). DBoW3.
URL: *https://github.com/rmsalinas/DBow3*

Odolinski, R. & Teunissen, P. J. G. (2017). Low-cost, 4-system, precise GNSS positioning: a GPS, galileo, BDS and QZSS ionosphere-weighted RTK analysis, *Measurement Science and Technology* **28**(12): 125801.

Odolinski, R., Teunissen, P. J. G. & Odijk, D. (2015). Combined GPS + BDS for short to long baseline RTK positioning, *Measurement Science and Technology* **26**(4): 045801.

Oliva, A. & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope, *International Journal of Computer Vision* **42**(3): 145–175.

OpenCV (2018). Opencv 4.0.
URL: *https://opencv.org/opencv-4-0/*

Paziewski, J., Sieradzki, R. & Baryla, R. (2018). Multi-GNSS high-rate RTK, PPP and novel direct phase observation processing method: application to precise dynamic displacement detection, *Measurement Science and Technology* **29**(3): 035002.

Realini, E. & Reguzzoni, M. (2013). goGPS: open source software for enhancing the accuracy of low-cost receivers by single-frequency relative kinematic positioning, *Measurement Science and Technology* **24**(11): 115010.

Rublee, E., Rabaud, V., Konolige, K. & Bradski, G. (2011). ORB: An efficient alternative to sift or surf, *2011 International Conference on Computer Vision*, pp. 2564–2571.

Schinazi, V. R., Thrash, T. & Chebat, D.-R. (2016). Spatial navigation by congenitally blind individuals, *Wiley Interdisciplinary Reviews: Cognitive Science* **7**(1): 37–58.

Song, T. (2014). LibGIST.
URL: *https://github.com/whu-tgsong/LibGIST*

Sünderhauf, N., Neubert, P. & Protzel, P. (2013). Are we there yet? challenging seqslam on a 3000 km journey across all four seasons, *IEEE International Conference on Robotics and Automation (ICRA) 2013*, Proc. of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA), Karlsruhe, Germany.

Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B. & Milford, M. (2015). On the performance of ConvNet features for place recognition, *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4297–4304.

Talbot, B., Garg, S. & Milford, M. (2018). OpenSeqSLAM2.0: An open source toolbox for visual place recognition under changing conditions, *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7758–7765.

Torralba, Murphy, Freeman & Rubin (2003). Context-based vision system for place and object recognition, *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 273–280 vol.1.

Vysotska, O., Naseer, T., Spinello, L., Burgard, W. & Stachniss, C. (2015). Efficient and effective matching of image sequences under substantial appearance changes exploiting GPS priors, *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2774–2779.

Vysotska, O. & Stachniss, C. (2017). Relocalization under substantial appearance changes using hashing, *Proc. Int. Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. Workshop Planning, Perception Navig. Intell. Veh.*

Yang, X. & Cheng, K. T. (2014). Local difference binary for ultrafast and distinctive feature description, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(1): 188–194.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A. & Torralba, A. (2017). Places: A 10 million image database for scene recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* .

The authors have confirmed that any identifiable participants in this study have given their consent for publication.